Summary

This is the complete summary of cleaning “**Netflix_titles_Dataset**” here are the steps that are followed to clean this dataset.

Tools

- Python (Jupyter Notebook)
- MS Excel

Steps

Python(Jupyter Notebook)

- import pandas as pd
  import numpy as np

- Imported file in df(variable) using this code **df = pd.read_csv(r"D:\Elevate Labs Internship\Day 01\netflix_titles.csv").**

- Get the overview of the dataset by using df.info().

- Get first 5 values of each column of the dataset using **df.head().**

- Using this df.isnull().sum() get the all the null values.

```
show_id          0
type             0
title            0
director      2634
cast           825
country        831
date_added      10
release_year     0
rating           4
duration         3
listed_in        0
description      0
```

- The null values in the Director, Cast, and Country columns have been replaced with 'Unknown'. This is because some movies and series do not provide information about their cast or director, and films can be produced in different countries. Keeping these values as NaN would be considered incomplete data.

```
df["cast"] = df["cast"].fillna("Unknown")
df["director"] = df["director"].fillna("Unknown")
df["country"] = df["country"].fillna("Unknown")
df.isnull().sum()
```

- `df["rating"].unique()` to get the unique values of rating column.

  ```
  array(['PG-13', 'TV-MA', 'PG', 'TV-14', 'TV-PG', 'TV-Y', 'TV-Y7', 'R','TV
  -G', 'G', 'NC-17', '74 min', '84 min', '66 min', 'NR', nan,'TV-Y7-FV', 'U
  R'], dtype=object)
  ```

  The rating column has some garbage values like 74 min, 84 min and 66 min to remove garbage values to null i used below code.

  ```
  df["rating"] = df["rating"].replace(["74 min","84 min","66 min"],np.nan)
  df["rating"].unique()
  df.isnull().sum()
  ```

  Now the value null values are like this

  ```
  show_id        0
  type           0
  title          0
  director       0
  cast           0
  country        0
  date_added     10
  release_year   0
  rating         7
  duration       3
  listed_in      0
  description    0
  ```

  Rating and duration nan values are not replaced with unknown, mean, median and mode because these values are irreplaceable.

- Replace the unnecessary space " " from the date_added column

  ```
  df['date_added'] = df['date_added'].str.replace(r'^\s', '', regex=True)
  ```

- Saved the cleaned dataset as a new file in the same location.

  ```
  df.to_csv(r"D:\Elevate Labs Internship\Day 01\netflix_titles2.csv", index=False)
  ```

MS Excel

- Make column header bolder.

- Arrange all the value in center.